Automated Probabilistic Analysis of Air Traffic Control Communications

Randy Jensen[1], Richard Stottler[2]
Stottler Henke Associates, Inc., San Mateo, CA 94404

and

Bonnie Schwartz[3]
Air Force Research Labs, Wright Patterson AFB, OH 45433

Initiatives to integrate autonomous Unmanned Aerial Vehicles (UAVs) with regular airport operations require automated onboard situational awareness to maintain safety at all times. More specifically, this requires the capability to sense, interpret, and predict what other aircraft are doing, based on the same incoming data that are available to a human pilot. This includes not only baseline knowledge of the airport layout, operational practices and landmarks, but also an ability to interpret radio communications with Air Traffic Control (ATC) and correlate them with observable movements and positions of other aircraft. This analysis informs an autonomous UAV's control mechanisms which ultimately regulate its kinetic behavior at the airport. As with any operational domain governed by human actions and control, there are many inherent challenges in interpreting ATC communications – a noisy data stream not only in terms of signal quality, but more significantly in the range of human deviations from the strictest procedures. This makes the analysis a natural application for Artificial Intelligence techniques, where the goal is to support automated reasoning that mimics a human pilot's decision processes. This paper provides a detailed discussion of a probabilistic reasoning approach using Bayesian Networks to classify ATC communications and synthesize them with baseline knowledge of an airport and produce real-time hypotheses about the states and trajectories of other aircraft. This provides a key component for automated situational awareness, which also requires correlation with sensor data, and ultimately a functional set of behaviors to act accordingly, although these latter capabilities are beyond the scope of this paper. The probabilistic communications analysis methodology is described, along with testing results using a real-world sample data set annotated for ground truth, to evaluate performance.

Nomenclature

ASR = Automated Speech Recognition
ATC = Air Traffic Control
BN = Bayesian Network
FAA = Federal Aviation Administration
UAV = Unmanned Aerial Vehicle

I. Introduction

INITIATIVES to integrate autonomous Unmanned Aerial Vehicles (UAVs) with regular airport operations require automated onboard situational awareness to maintain safety at all times. More specifically, this requires the capability to sense, interpret, and predict what other aircraft are doing, based on the same incoming data that are

[1] Group Manager, 951 Mariner's Island Blvd., Suite 360, San Mateo, CA 94404, AIAA Contributor.
[2] President, 951 Mariner's Island Blvd., Suite 360, San Mateo, CA 94404, AIAA Member.
[3] USAF AFMC AFRL/RQQC, 2130 8th Street, Bldg. 45 Rm. 283A, Wright-Patterson AFB, OH, 45433-7542, AIAA Contributor.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
AUG 2013		**00-00-2013 to 00-00-2013**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Automated Probabilistic Analysis of Air Traffic Control Communications	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Stottler Henke Associates, Inc,951 Mariner's Island Blvd., Suite 360,San Mateo,CA,94404**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of the Infotech@Aerospace 2013 Conference, American Institute of Aeronautics and Astronautics, 19-22 Aug, Boston, MA.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**9**	

available to a human pilot. This includes not only baseline knowledge of the airport layout, operational practices and landmarks, but also an ability to interpret radio communications with Air Traffic Control (ATC) and correlate them with observable movements and positions of other aircraft. This analysis informs an autonomous UAV's control mechanisms which ultimately regulate its kinetic behavior at the airport. As with any operational domain governed by human actions and control, there are many inherent challenges in interpreting ATC communications – a noisy data stream not only in terms of signal quality, but more significantly in the range of human deviations from the strictest procedures. This makes the analysis a natural application for Artificial Intelligence techniques, where the goal is to support automated reasoning that mimics a human pilot's decision processes.

Aside from the speech recognition task where humans naturally excel, other sensemaking tasks similarly involve methods that humans routinely employ consciously or subconsciously, such as the consideration of context, comparison with past experience, and recognition of distinct individuals participating in communications. While some natural reasoning methods are difficult to model in a computer-based system, others make use of patterns that can be encoded as heuristics in evidence functions for reasoning that takes place after the speech recognition task. This is the focus of this paper, describing a probabilistic reasoning approach using Bayesian Networks to classify parsed ATC communications and synthesize them with baseline knowledge of an airport to produce real-time hypotheses about the states and trajectories of other aircraft. This provides a key component for automated situational awareness, which also requires correlation with sensor data, and ultimately a functional set of behaviors to act accordingly, although these latter capabilities are beyond the scope of this paper. The probabilistic communications analysis methodology is described, along with testing results using a real-world sample data set annotated for ground truth, to evaluate performance.

II. Background

The task of automating the analysis of ATC communications to support autonomous UAV operations involves three major subtasks for each transmission.

- **Attribute speaker**. There are two elements to this attribution. First, the analysis must determine which aircraft a transmission concerns. The second goal is to resolve the finer question of whether the speaker is the pilot or the controller.
- **Classify the type of communication and associated procedural step**. When a transmission occurs, the classification objective is to identify what it implies about the procedural state of an aircraft, in terms of what it is doing and possibly what it is about to do. This entails identifying not only the associated operational procedure such as taxiing in preparation for departure, but also the nature of the transmission in terms of the communication sequence associated with the procedure. For example, many procedures involve communication sequences structured in a similar way, with transmissions for a request, approval, and confirmation. Thus an individual transmission might be classified as the ATC approval for the taxi procedure, and attributed to a specific aircraft. Not all transmissions are associated with any significant procedural step, and therefore a secondary objective of the classification task is also to minimize false positives. In other words, it is also important to minimize results where an insignificant transmission is incorrectly classified as something procedurally significant.
- **Extract content**. Given a classification for a transmission, it is also necessary to extract any operationally relevant content that can be used for the reasoning about current and future states of aircraft. For example, in a transmission containing a taxi route, the route information must be extracted and passed to the modeling component used by the autonomous UAV control mechanisms.

The overall objective is to use the outputs of these analysis tasks to correlate spoken information about aircraft call signs, types, and locations with ATC instructions for upcoming activity, and then with physical tracking data coming from sensors.

The communications analysis tasks depend heavily on a structured model for procedural operations in the terminal area of an airport, which can be grouped into a relatively small set of categories. The primary categories are taxi and ground movement, spacing and sequencing, departure, and arrival. Formal conventions for these procedures are outlined in Federal Aviation Administration (FAA) documents, providing a clearly structured baseline for typical operations, as well as for the analytical task of interpreting operations. However, the real-world operating environment still involves a great deal of ambiguity and uncertainty from the perspective of a computer-based analytical system, due to the kinds of challenges summarized in the following Table 1.

Table 1. Challenges in automated ATC communications analysis.

Analytical Challenge	Details
Speech recognition performance	Limitations in automated speech recognition (ASR) software performance mean that a certain degree of parsing inaccuracy should always be expected.
Radio signal quality	Variations in radio signal quality can further degrade the reliability of ASR results.
Procedural deviations	Deviations from formal FAA procedures may be customary in practice at different airports, particularly at smaller regional airports.
Unknown speaker	Attribution of the speaker in a radio transmission comes only from the content of the transmission itself through verbal self-identification, as opposed to any direct means in the data stream to explicitly identify the speaker.
Omitted call signs	Even under direct FAA procedures, it is allowable to omit call signs in radio transmissions under certain conditions. This is reasonable for humans who naturally understand context within a dialog, and also recognize individual voices. But these same capabilities don't come naturally to an automated system.
Overloaded terms	Interpretation is complicated by the repeated appearance of overloaded terms, especially those that are part of the radiotelephony phonetic alphabet, such as "delta." For example, consider the following transmission collected at the San Jose International Airport. "*avantair one thirty eight san jose ground verify information delta runway three zero left intersection delta taxi via victor delta cross runway two niner*" In this transmission, "delta" appears three times: once as the designator for automatic terminal information service status ("information delta") and twice in the description of a ground taxi route, describing a runway intersection and a taxiway. In other cases, "delta" may also be used in reference to an airline carrier, or part of a tail number identifier. In order to correctly interpret the transmission, these overloaded terms must be correctly associated with the right meanings in the context where they appear, or at least tagged with different possible associations for different competing hypotheses that will be tried.
Human error	Mistakes also occur in human communications, as in the following examples. "*skywest correction southwest eleven fifty nine wheels up at four niner*" "*southwest nine uh twenty nine twenty one taxi*" While there are formal ways to indicate an error, such as the keyword "correction," there are also informal methods that humans easily understand but strict parsing rules may not.

The challenges above make a clear case for a probabilistic approach that attempts to mimic some of the thought processes that humans would also use to interpret ATC communications. In contrast to more direct rule-based methods, Bayesian Networks have proven effective at handling noisy and uncertain input data in many applications. A Bayesian Network (BN) is a graphical modeling structure used to represent acyclical connections between variables in order to predict relationships under conditions where observed inputs may be lacking for some of the variables.[1] In practical terms applied to this domain, each of the challenges outlined above translates to an impact on either the fidelity or availability of a variable used to represent evidence for the proper classification of an ATC radio transmission. For example, an omitted call sign means that no information is available about the input variable associated with matching a transmission with a particular classification and known aircraft. While other rule-based models will often generate inaccurate predictions when an input variable is missing, BNs are still able to perform effective reasoning because of the fact that they encode the dependencies between variables.

More generally in the ATC communications analysis task, the goal is to extract the maximum practical useable information from a transmission, both in terms of its own classification and content as well as its potential implications about previous or future transmissions. The following example illustrates some of the desired contribution from Bayesian reasoning, in a situation where ASR results are suboptimal. Two transmissions are parsed, with one immediately following the other. In the first transmission, several words (or phonetic combinations) receive low confidence scores, and are shown here with the placeholder "[unknown]."

(1) *"[unknown] [unknown] runway three [unknown] taxi via foxtrot [unknown]"*
(2) *"foxtrot yankee to runway three zero right southwest six two two"*

This overall pattern fits with a taxi instruction and confirmation, which means it's likely that (1) is from the ATC ground controller, (2) is from the pilot of Southwest 622, and both reference the same taxi route. Because of the uncertainty of the parse content in (1), it is almost unusable for anything other than a probabilistic approach, because of the number of input variables that are missing. However, a BN model provides a formalized way to produce hypotheses for this transmission which make best use of available evidence, and perhaps more importantly, establish a grounds for more definitive reasoning about transmission (2) which follows immediately after.

Specifically in transmission (1) above, the call sign is absent, it's unclear whether the speaker is a pilot or controller, and there are only partial snippets of a taxi route description. However, the presence of the keywords "taxi via" provide useful evidence to narrow down the relevant procedure, as well as the reference to "runway three" for this airport, where there are runways named "three zero right" and "three zero left." Since the next transmission (2) is structured like a pilot communication, with content elements matching the partial elements found in (1), this establishes two kinds of evidence for input variables needed to classify (2) as a taxi confirmation. First, the content matches, and second, there is an apparent relationship with the preceding transmission where both were part of the same dialog. A human trying to interpret the same pair of transmissions would consider the same kinds of evidence, using not only the direct content of each transmission, but also the linkages between transmissions, and other forms of context.

III. Probabilistic Reasoning Approach

As described above, the objective of automated analysis in this domain is to produce classification hypotheses for parsed ATC radio communications, where each hypothesis pairs a type of dialog event with a specific tracked object (i.e., a specific aircraft). A single transmission is considered a dialog event, and a collection of adjacent dialog events often forms a coherent dialog. So an individual dialog event may result in a significant classification that causes state models to transition, or it may just have a role as an element of a dialog in progress. Therefore, in the analysis of a single dialog event, the reasoning system must perform separate tests for the same classification, paired with different objects (e.g., taxi confirmation for Southwest 622, versus taxi confirmation for United 9984, etc.). In order to manage this process in execution, the BNs are constructed as templates, where each template is associated with a single dialog event classification (such as taxi confirmation). These templates are then instantiated for each object tracked, as illustrated in Figure 1 below.

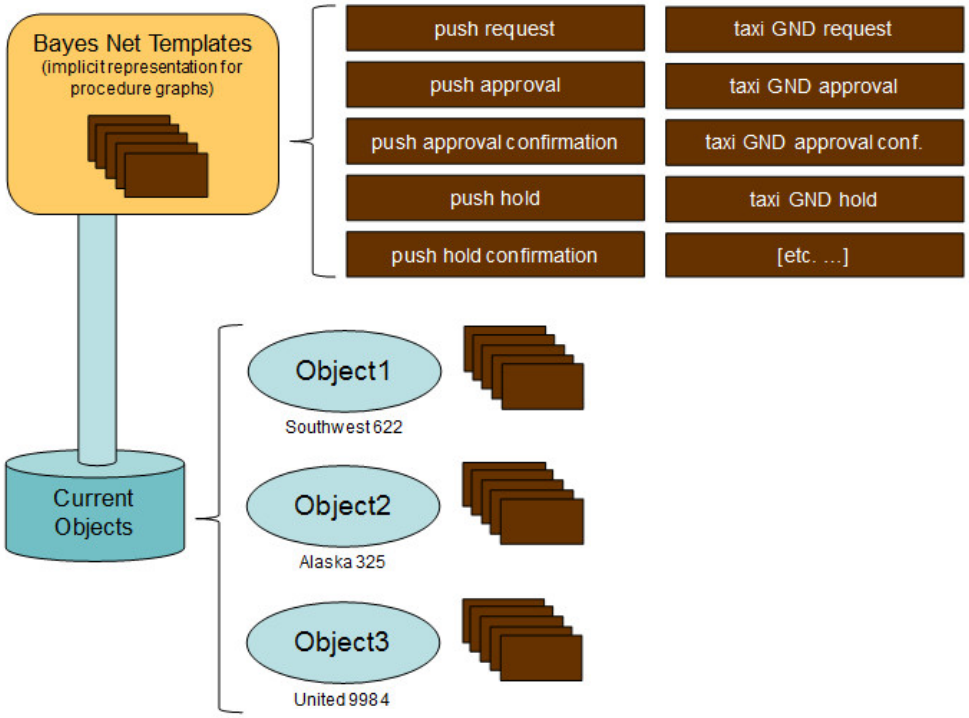

Figure 1. BN templates and execution. *For each object tracked, a set of BNs is instantiated from the templates.*

Thus as the repository of currently tracked objects grows, the number of BNs that may be tried with each dialog event grows as well. However, the computational demands remain manageable because of the relatively small number of tracked aircraft objects in the terminal area of an airport at any given time. Also, the BNs themselves are designed to be as simple as possible, both for the purposes of effective knowledge engineering in weighting input variables, and for computational speed.

In order to produce a simple BN design, we defined a generalized set of evidence categories that can be used as input variables in the BN graph structures. These evidence categories represent an attempt at thorough coverage of the dialog event features that contribute to reasoning about a classification hypothesis, whether this reasoning is performed by a human or an automated system. Table 2 below describes the evidence categories.

Table 2. Evidence categories used as BN input variables.

Evidence Category	Details
Identity Consistency (ID)	Is the parsed aircraft call sign (if any) consistent with the identifiers for the aircraft that this dialog event is matching with?
Word Contents (WC)	Does the transmission parse contain the expected words for this kind of dialog event?
Word Ordering (Order)	Does the word ordering in the transmission parse match this kind of dialog event? The word ordering test is primarily concerned with determining whether the speaker is a pilot or controller. For example, a pilot and controller may both refer to the aircraft call sign and the ATC unit, but the ordering within the transmission often (but not always) implies which one is speaking.
Dialog Pattern (Dialog)	For dialog event types that are typically expected to appear within a dialog sequence (e.g., request, approval, confirmation), does this transmission match a pattern with preceding dialog events in terms of timing and sequencing that is consistent? For example, as an input variable inside the BN for a taxi confirmation classification hypothesis, this evidence category involves an assessment of the preceding one or two dialog events, to see if they were classified as a taxi request and taxi instruction, respectively. The Dialog Pattern category does not apply with all dialog event types, because those that are naturally the first in a dialog sequence do not have any specific expected pattern of preceding dialog events to match with.
History Consistency (History)	For dialog event types associated with a procedural state that implies a history of prior states, is this consistent with the history of activity (and timing) that has been established for the aircraft that this dialog event is matching with? For example, as an input variable inside the BN for a taxi confirmation classification hypothesis with the object Southwest 622, this evidence category involves checking the existing state hypotheses for that object to determine if they are consistent with a taxi confirmation.

All classification BN templates are constructed with evidence nodes for these categories, so in practice this means that the graphical representations have at most 6 nodes (5 for these forms of evidence and one for the classification itself). Each node corresponds to a continuous variable with an associated evidence function that dynamically calculates a value based on the features of the current transmission. Given the uncertainty in the data stream, all evidence functions return scored values, also known as virtual evidence. For example, the evidence function for Identity Consistency doesn't return a simple true or false value, but also takes into account the confidence score on the parse where the aircraft call sign was found, and the quality of the match. This allows for uncertain or imperfect input data to contribute to the reasoning process in each of the evidence categories.

Although the classification BN templates are constructed with nodes that reuse roughly the same set of evidence categories across all the different dialog event types, the differences lie in the evidence functions and the conditional probabilities associated with the nodes in different templates. The following example illustrates how this is constructed in practice. An example dialog sequence is shown in Table 3, using three transmissions collected from real operations at San Jose International Airport. These three transmissions occurred immediately one after the other, in a standard dialog sequence for initiating the pushback procedure from a gate.

Table 3. Example dialog sequence.

#	Time	Parse	Ground Truth
(3)	50:46	*ground southwest four thirteen push door nineteen with uh foxtrot*	Push Request (SW-413 requesting push from gate 19)
(4)	50:52	*south four thirteen san jose ground push approved*	Push Approval (KSJC Ground giving push approval)
(5)	50:54	*thanks*	Push Approval Confirmation (SW-413 confirming the push approval)

In the table above, each transmission is identified by number for reference, with the start time when the transmission occurred, the parse results from speech recognition, and a ground truth column reflecting what the actual dialog event type is. The analysis objective is to produce classifications that match the ground truth. For the sake of simplicity in a discussion of the probabilistic reasoning methods, this example ignores the question of speech recognition performance and assumes that the results are perfect (i.e., the parses match exactly what was actually spoken, with high confidence scores). Even in this nearly ideal condition, there are still elements that make the classification problem challenging. In transmission (3), the call sign as spoken is not an ideal match (the abbreviated "south" instead of "southwest"). And transmission (4) gives almost nothing useable in terms of content, yet it is an important part of the procedural analysis to determine that SW-413 acknowledged the push approval.

The following discussion steps through the evidence function results for the relevant classification BNs. Although in practice each transmission in the example above is tested with many BNs, this discussion focuses only on those corresponding to the ground truth classifications above. Figure 2 shows the evidence function results for transmission (3) when it is tested with the BN for the Push Request classification paired with the known aircraft object, Southwest flight 413. In this case, all relevant evidence functions return their maximum values. The parsed call sign is a clean match with Southwest 413. The WC evidence function looks for specific keywords like "push" and a reference to a gate, which are also both matched. The Order function finds "ground" before the pilot call sign, which is consistent with the transmission being from the pilot. The Dialog function is not applicable in this BN because it is associated with a dialog event that typically appears first in a sequence (the push request). The History function finds that a push request is consistent with the current state of Southwest 413, which had previously checked in but not initiated any other procedures. With the values returned by these evidence functions used as input to the BN, the resulting score for this classification hypothesis is 0.99 (very high).

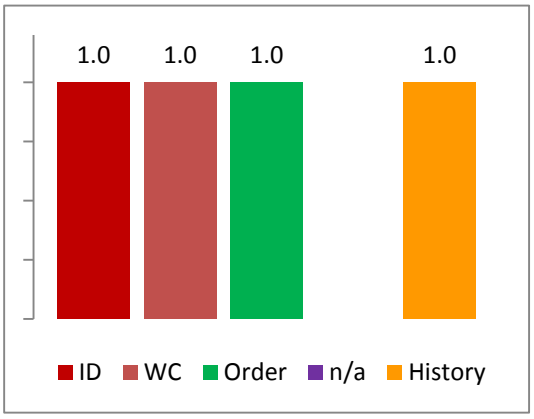

Figure 2. Evidence function results for a Push Request classification on transmission (3).

Next, Figure 3 shows the evidence function results for transmission (4) when it is tested with the BN for the Push Approval classification paired with Southwest flight 413. Once again, most of the evidence functions return top scores. This is an example where the WC function looks for different content depending on the dialog event type. Because this BN is associated with the Push Approval dialog event classification, the WC function also looks for keywords like "approved," but these keywords would not be expected in a Push Request. In this case, the parsed call sign is "south four thirteen", which is not a perfect match with the object

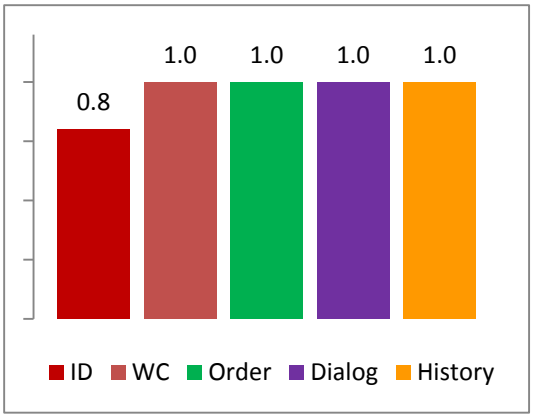

Figure 3. Evidence function results for a Push Approval classification on transmission (4).

Southwest 413 that is paired with this version of the BN template. Therefore the ID function returns a reduced value in this case, although it benefits from the same flight number and the fact that "south" is close to "southwest". The Dialog function is applicable and significant in this BN, because a Push Approval is expected to follow a Push Request. And in this case it returns a top value because the immediately preceding transmission received a Push Request classification as the top scoring hypothesis. With the values collectively returned by these evidence functions used as input to the Push Approval BN, the resulting score for this classification hypothesis is 0.998 (very high).

Unlike the preceding two transmissions in the example dialog, transmission (5) provides very little content since it is a one-word response ("*thanks*"). Figure 4 shows the evidence function results when tested with the BN for the Push Approval Confirmation classification, paired with Southwest 413. In this case, there is no match with the ID, expected word contents, or the word ordering for a radio transmission from a pilot. However, this is very common for a Push Approval Confirmation, and thus the BN template for this dialog event type establishes conditional probabilities for these variables that decrease their significance in this case. This is very different from the conditional probabilities established for the same categories of evidence in the BN templates for the preceding two dialog event types. For example, the word contents ("push approved") and ID match are very important in a Push Approval classification, and so they are weighted accordingly in the BN. In this case, for the Push Approval Confirmation, the Dialog evidence function is actually the most significant, because it establishes that the transmission must immediately follow a Push Approval. The top scoring hypothesis for the preceding transmission was a Push Approval, paired with the object Southwest 413. Therefore the Push Approval Confirmation BN paired with Southwest 413 is the only hypothesis that can get a high value for the Dialog evidence function. As a result, because of the conditional probabilities in the BN template, even with low values for transmission (5) in three of the evidence categories, the BN score for the Push Approval Confirmation hypothesis is 0.862, the top scoring classification found.

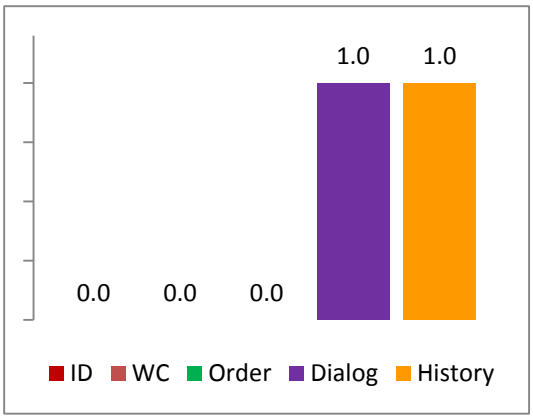

Figure 4. Evidence function results for a Push Approval Confirmation classification on transmission (5).

The general procedure for generating classification hypotheses and managing them over time is impacted by the use of evidence functions looking for Dialog patterns across transmissions. As a policy, it is useful to retain competing hypotheses because a low scoring hypothesis may have been missing one or more input variables due to uncertainty. Referring back to the earlier example with transmissions (1) and (2), it is the Dialog evidence function that helps classify this pair as a taxi instruction and confirmation. Given the missing input data for transmission (1), most likely this would receive many classification hypotheses with scores that are low, but above threshold. If the policy were to discard all but a set number of hypotheses (e.g., the top three), the outcome could be that the Dialog evidence function would fail to find a dialog pattern for transmission (2). However, it is most likely impractical to use a large number of competing hypotheses in the reasoning that triggers state changes for the purpose of the higher level end goal of affecting autonomous UAV control decisions, because of the expansion of the decision space. Thus the preservation of competing hypotheses largely serves the internal reasoning purposes of the communications analysis component, but these remain layered at a lower level than the exported states.

IV. Initial Experimentation and Results

The initial implementation of the probabilistic reasoning system includes a set of BN classification templates and their constituent evidence functions, developed for a subset of the terminal area procedures, specifically those related to taxi and ground movement. The ATC communications for these procedures are all conducted on the Ground channel, and so in order to test the initial implementation, three data sets were prepared by collecting Ground channel radio communications from the San Jose International Airport. The data sets consist of a total of 194 transmissions, manually transcribed into text. At least for initial experimentation, the objective was to isolate the classification task from the speech recognition task. Both tasks are non-trivial, but the latter can introduce variables that make it difficult to quantify the effectiveness of a probabilistic classification approach.

ATC radio communications routinely include both significant dialog events and also insignificant miscellaneous discourse. For the purposes of this application, a significant dialog event is any transmission with a meaningful contribution to an understanding of the aircraft's current or future procedural state. An insignificant transmission is the opposite, with little relationship to aircraft procedures. The variations that can be considered insignificant are virtually unlimited, such as the following transmission on the Ground channel, "*hey over there where they tore all that stuff down on the other end of the terminal here, what's that going to be parking or an extended terminal later?*"

The experimental data sets were constructed with a ground truth field for every transmission, which contains the appropriate dialog event classification, if any, for the transmission. All significant dialog events in our test data sets have a ground truth classification which corresponds to a BN classification template. All insignificant transmissions have a blank value for the ground truth classification. The experimentation objective is to compare automated analysis results with ground truth, which means generating classification hypotheses with scores and then comparing with expected classification outcomes as specified in the ground truth field. Significant dialog events should ideally produce classification hypotheses where the top scoring hypothesis is above an acceptable threshold and matches with the ground truth classification. Insignificant transmissions should ideally produce zero classification hypotheses above threshold. The experimental data sets are not filtered to remove insignificant transmissions, because it is important to assess performance in terms of both positive and negative cases. Approximately 60% of the transmissions in the data sets are significant dialog events, which means a relatively high proportion of 40% should receive no classification at all.

In order to effectively interpret the results of comparing automated classification results with ground truth, experiments were designed to characterize the quality of the match using five different exclusive grades.

- **True Positive**. The top scoring classification hypothesis matches ground truth and is above threshold.
- **True Negative**. The ground truth is blank for this transmission in the data set, and (correctly) no classification hypotheses are generated with above threshold scores.
- **Acceptable Positive**. One or more of the generated classification hypotheses matches ground truth and is above threshold, but the top scoring hypothesis does not match ground truth.
- **False Positive**. The ground truth is blank for this transmission in the data set, but (incorrectly) at least one classification hypothesis was generated with an above threshold score.
- **False Negative**. The transmission has a ground truth classification, but there are no generated classification hypotheses with above threshold scores, whether matching ground truth or not.

Experiments were conducted with three data sets collected at different times, with results counted for each possible kind of match. Notably, there were no instances of Acceptable Positive results with any of the data sets, so this is omitted from the results. Table 4 below shows test results with each data set, and a combined total in the last row.

Table 4. Classification results with test data sets.

Data Set	True Positive	True Negative	False Positive	False Negative
1 (30 min, 50 transmissions)	36 72%	14 28%	0 0%	0 0%
2 (30 min, 38 transmissions)	24 63%	14 37%	0 0%	0 0%
3 (90 min, 106 transmissions)	57 54%	41 39%	2 2%	6 6%
Combined (150 min, 194 transmissions)	117 60%	69 36%	2 1%	6 3%

The True categories represent success, so these results reflect a combined 96% success rate on the full collection of 194 transmissions. Given the small number of False results, further investigation of the failure conditions show that some of these are extremely challenging examples. For example, one of the two False Positive results comes from the following transmission.

(6) *"okay get confusion between gate twenty one and the other one pushing so the aircraft pushing off of gate twenty four say again your numbers"*

Transmission (6) receives an incorrect classification hypothesis as a Push Request, with an above threshold score. The ground truth field for this transmission is blank, because it is a miscellaneous communication from an ATC controller trying to resolve some confusion, so it should not be classified as any of the significant dialog events. However, the Word Contents (WC) evidence function for the Push Request classification BN finds keywords such as "pushing" and gate numbers, which in this case are enough to generate an above threshold score for the entire hypothesis. It is worth noting that no other classification hypotheses related to the pushback procedures are generated with acceptable scores. For example, the Push Approval hypothesis fails because its WC evidence function is looking both for words like "push" or "pushing" and also for "approved," which does not appear in this transmission. Ultimately in the experiment with this data set, the False Positive result with a Push Request hypothesis for transmission (6) has no impact on the object state tracking, because there is no complete push dialog. Without a subsequent Push Approval and Push Approval Confirmation, there is no trigger a state change for any tracked object.

Investigation of the False Negative results from these experiments reveals that all are caused by object attribution problems. In many of these cases, we expect performance to improve with refinements in the flexibility of the matching capability in the ID evidence function. For example, two of the six False Negative cases occur because an aircraft with the actual call sign of "Southwest 2989" was identified as "Southwest flight 989." This leads to a matching failure which impacts both transmissions in terms of the ID and Dialog evidence functions. Although there are many ways in which call signs are abbreviated or modified from time to time in verbal communications, many of these kinds of variations can be handled with rules to accommodate partial matches. This kind of flexibility is currently being added to the existing capability, and will generalize to all dialog events that involve matching with a call sign.

V. Conclusion

The 96% success rate in initial experiments is very promising for the feasibility of applying a Bayesian probabilistic approach to classifying ATC communications. The factors that were controlled for this testing are the natural next steps to introduce into the problem space. For example, the use of human transcribed text as the input source controlled for variations in speech recognition performance, but this is one of the major areas of uncertainty that the probabilistic approach will attempt to resolve. Thus a natural direction for future research is to incorporate ASR outputs into the testing data sets, evaluate performance, and attempt to adapt the conditional probabilities and evidence functions in the BN templates accordingly to maintain a high level of success. While the addition of realistic ASR input is likely to have a negative impact on the overall classification success rate, there are other factors that may exert influence in the opposite direction. Perhaps the most significant secondary form of contextual input that human pilots use in reasoning about the movements of other aircraft in the terminal area is the observation and correlation of visible aircraft with the communications they hear. Thus in the automated reasoning system, the integration of sensor data will provide a valuable independent source of information to help confirm or reject classification hypotheses, especially where the communications stream itself carries a great deal of uncertainty.

Acknowledgments

The effort described in this paper was sponsored by the US Air Force Research Laboratory. The views expressed in this paper are those of the authors and do not necessarily represent the official policy or position of the Department of Defense, the US Air Force, or the US Air Force Research Laboratory.

References

[1]Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Iinference* Morgan Kaufmann, San Francisco, 1988.